FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 3 No. 1, March, 2019, pp 56 - 64

IMPROVED ELECTRONIC MAIL CLASSIFICATION USING HYBRIDIZED ROOT WORD EXTRACTIONS

Okunade, O. A.

Department of Computer Science, Faculty of Sciences, National Open University of Nigeria, Abuja, Nigeria.

Correspondence Author's Email: aokunade@noun.edu.ng

ABSTRACT

Content based spam filter prevents spam mail from successful delivery to the targeted host using Bayesian probability approach. Unfortunately, spammers deceived content based filters by coming up with sophisticated means of circumventing detective pattern of developed content filters, manipulating and rearranging spam mail suspicious terms/content to fool such filters, since content based spam filters only work effectively, if the suspicious terms are lexically and grammatically correct. However, this paper proposes word stemming combined with Bayesian probability approach to regain spam-free inbox in the electronic mail infrastructure. The hybridized technique was used to detect modified suspicious terms by examining the base root of the misspelled or modified manipulated suspicious words/terms and reconverting them to the correct token or near correct token and examine as such. The implementation of the algorithm when tested with direct and manipulated spam mail content was able to successfully identified spam mail with manipulated suspicious terms and 99% of the tested known manipulated suspicious terms spam mail were identified and classified as spam. However manipulated spam mail is of no effect in hybridized word stemming combined with Bayesian probability spam filter approach. The algorithm is effective, accurate, prevent false classification and negate spammer's innovation.

***Keywords*:** Spam, Ham, Email, Suspicious terms, Stemming, Filter, Spammer.

INTRODUCTION

In recent time, developments in Internet communication with applications such as World Wide Web (www), precisely electronic mail (email) increases the usefulness and availability of tremendous services to users all over the world. Despite the availability and usefulness of Internet, it could be hazardous due to its negative exploitation (Andrej, Gordon, Bogdan, Thomas and Blaˇz, 2006). Email is undoubtedly one of the Internet's killer applications, though it satisfies the basic human need for communication and has become a critical mission in every organization. But can be frustrating, time wasting and, devastating which resulted into various types of lost and high consumption of (data storage, bandwidth, finances and power) (Priyanka & Prashanthi, 2015 and Reshma & Dhanya, 2017). Dealing and classification of spam is a very difficult task, a single model classification cannot tackle the problem (Mrutyunjaya, Ajith & Manas, 2011). Due to new spam that are constantly evolving and often actively tailored not to be detected (Rekha and Sandeep, 2014). According to the study by Aladdin Knowledge Systems (2011) in Omar, Ashraf and Ramadan (2012) it was estimated that over 70% of today's business emails are spam and if proper consideration does not take to the effect, it could escalate beyond easy control. Hedieh, Golazin & Fatemeh (2016) spam varies across the region, for instance, in North America less than 1% of SMS messages were spam in 2010, while in parts of Asia up to 30% of messages were spam messages. In China and during 2008, the number of daily sent messages was 1.9 billion, and China's mobile phone users received an average of 10.35 spam messages per week3. According to Sarah, Mark and Derek (2012) SMS spam contributing to 20-30% of all SMS traffic in China and India. Spam can be categorized into the various listed categories according to (Thamarai, Hamid and Alaa, 2010) based on Ferris Research (2009):

1. Health: Examples of this are forge pharmaceuticals.
2. Promotional products: Examples of this are forge fashion items (for example, watches, cloths, costumes and so on);
3. Adult content: Examples of this are pornography and prostitution shows and promotion
4. Financial and refinancing: Examples of this are stock kiting, tax solutions, loan packages.
5. Phishing and other fraud: Examples of this are "Nigerian 419" and "Spanish Prisoner".
6. Malware and viruses: Examples of this are Trojan horses attempting to infect PC with malware.
7. Education: Examples of this are forge online diploma.
8. Marketing: Examples of this are direct marketing material, sexual enhancement products.
9. Political: Examples of this are political votes.

Likewise, there are various availability of content based spam filtering techniques that can be use to separate spam from important mails according to (Blanzieri & Bryl, 2008) and Zhang Zhu & Yao, 2004 in Amol, Prashant & Anil, 2013); such as: Naïve Bayesian classification, Support Vector Machine, K Nearest Neighbor, Neural Networks and so on.

There are several billions of emails delivered daily connecting people around the globe, the majority of all emails circulating on the Internet are unsolicited bulk emails called spam (Albercht, 2006). Spam presently contributes about 90% of all email on the Internet (Out-law News, 2006). The unwanted / unsolicited bulk email (UBE) or unwanted / unsolicited commercial email (UCE) are called spam according to (Jonathan, 2003 and Samir & Elsagheer, 2013). Day by day the amount of incoming spam increases and scammer attacks are becoming more of a threat to Internet community (Nazirova, 2011). Spam can also be defined as junk mail that are mostly advertisement material, it is a subset of electronic spam involving nearly identical messages sent to various recipients by email. Blank spam may also occur when a spammer forgets or otherwise fails to add the payload when setting up the spam according to (Anbazhagu, Praveen, Soundarapandian and Manoharan, 2014). Most commonly used varieties spam are advertising spam, blank spam, image spam, backscatter spam, social network spam, blog spam, forum spam and search engine spam. It is used for advertising products and services typically related to adult entertainment, quick money and other attractive merchandises according to (Cranor and LaMacchia,1998 and Kanich, Weaver, McCoy, Halvorson, Kreibich, Levchenko, Paxson, Voelker and Savage, 2011) in (Ja'far, Hossam, Khalid, Malek and Omar, 2015).

Spam messages can be quite harmless (Ham) or vice versa, to bring a potential threat (Siham, Wadeea, Ahmed and, Ibrahim, 2015). Many small companies that have legitimate business transaction also uses spam e-mail in order to advertise their legitimate products and services, but lesser compared to increased rate of malicious spam that adds another dimension to the adverse nature of spam email. This touches privacy and security of individuals and organizations according to (Ja'far, Hossam, Khalid, Malek and Omar, 2015). In computing, spamming is a criminal activities using social engineering techniques (Fight Cybercrime (November, 2008). Spamming is the abuse of any electronic communication medium to send unsolicited messages in bulk, the purpose of spam can be to make money off a product, spread viruses, chain letters, advertisement, political advocacy, fraud attempts or to get personal information from users. Spammers attempt to fraudulently acquire sensitive information, such as usernames, passwords and credit card details, by masquerading as a trustworthy entity in an electronic communication. eBay and PayPal are two of the most targeted companies, likewise online banks are common targets. Spamming is typically carried out by email or instant messaging and often directs users to give details at a website (Fight Cybercrime, 2008).

Ham is legitimate mails, it is used when referring to genuine email that is, the opposite of spam. Junk mail was already an issue in 1975 when John Postel wrote a "Request for Comments on the junk mail problem". The problem has been growing since then and now every email user knows what spam or junk mail is all about (Postel, 1975). To prevent spam from becoming email's killer application, a plethora of countermeasures have been proposed, for instance legal regulations, DNS-based attempts, content based, and a variety of solutions exploiting different spam filtering techniques. However, the content based filtering is one of the effective method used in tackling the spam, but yet the scammer sabotage the capability of the content based filter (Bayesian to be precisely). By introducing unnecessary special characters in between the suspicious terms/words and rearrangement of suspicious terms to defraud the filter. By writing terms that are lexically/ grammatically incorrect in order to deceive the content based filters, whereas meaningful to the readers. With this, scammers aim is being achieved, because those modified/rearranged words has meaning and understanding to the readers due to the fact that we do not read the letter(s) in the word one after the other but we read the word as a whole, for instance: database written as: 'dtaabase', 'datbaase', or 'dabatase', Viagra written as: 'Via*gra', 'Vi\gra!', 'V.i-a.g*r.a', '\/iagra', rich written as: 'r_i_c_h', 'r.i.c.h', 'r*ich', 'r1ch', 'r|cih', and so on. Users can still infer the correct meanings from those listed set of words, while the content filter cannot identify them as a suspicious words since it was trained with lexically corrected tokens. The topic area is very important, useful and desperately called for urgent intervention, to prevent further exploitation of scammer and reoccurrence of such manipulation in other areas of content based filter. The rest of this paper is structured as follows: Literature review discusses the similarities and differences between email spam and ham, and content-based spam current research focus area. While methodology and material discusses the method and approaches applied in tackling the issues with spam. Result and discussion displayed and analyzed the outcome of the research and conclusion concluded the paper.

MATERIALS AND METHODS

Word Stemming and Bayesian Probability Combined

The Bayesian probability formula was used based on its previous record of successful classification of e-mails. It provides information on the probability of an event occurring based on the probabilities of two or more independent evidentiary events. It has played a long role in accurate spam identification and is one of the reliable methods of spam identification. But yet, scammers can fool it by introducing semantically correct but lexically wrong words. However, the two techniques (word stemming and Bayesian probability combined) were brought together to make use of the successful record of email classification power of Bayesian probability. And cater for its deficiency using the word stemming to prevent further manipulation of suspicious words and expose its operational hide out.

Word stemming method can be used to scrutinize the main content of incoming mails, in order to match the identified suspicious terms against the particular domain suspicious key words stored in a database. Since as long as the first and the last letters of a word remained in place it does not matter what order the others are rearranged, most readers are able to

recognize the words as the original/real word (John, 2003). Stemming has been used as a technique to remove unwanted prefixes, affixes and suffixes in a word in order to generate its actual root word. In this paper, word stemming is combined with Bayesian probability to further improve email classification efficiency, by extracting the root content/meaning of each terms manipulated by the scammers to defraud the filter, for accurate classification of spam mails. This will create a highly efficient filtering scenario for instances where suspicious terms were manipulated.

Classification Using Root Based Extraction

When an email arrives through the Mail Transfer Agent (MTA), a new filtering process is initiated and pass the mail to the word root extractor algorithm that initiate the word Stemming Techniques operation. The algorithm sequence of operation during implementation is provided as follow:

Operation Sequence

(1) Counter initialize to value 1

(2) The Algorithm checks if the term to be checked/passed is less than or equal to the total mail terms (that is, counting the entire words one after the other (if term<=mail.net)) if true, the following operation will perform on the terms.

(3) The occurrence of any unwanted special characters is checked, if confirmed they will all be extracted. Such as special characters used to misspelled/manipulate/modified tokens (such as, $, /, \, |, =, !, @, #, %, ^, &, *, (,), <, >, ?, :, ”, ’, {, [, },] and so on).

(4) The algorithm check for the present of suspicious terms, by matching each terms one after the other against the list of suspicious terms present in the database table. If matched, the algorithm retrieve the particular term's Spamicity value stored in the database table (every suspicious terms in the database has its own assigned spamicity value). The value is then used to compute the Bayesian formula that contributes to the overall spamicity value of the entire e-mail.

(5) Finally, the algorithm rearranges the manipulated terms to its actual real term and calculate its spamicity value using the Bayesian Probability formula as follow:

$$P(a,b,c)=\frac{a*b*c}{a*b*c+[(1-a)*(1-b)*(1-c)]} \quad \text{..... equ 1}$$

a = is the first mail term, to be fund in the Suspicious terms table of the database,

b= is the second mail term, to be fund in the Suspicious term table of the database,

c= is the third mail term, to be fund in the Suspicious terms table of the database,

z = which is the last mail term, to be fund in the Suspicious term table of the database.

Based on the forgoing, if the (Probability of (a,b,c,…z) or (The end result calculated) <= 0.5 the recipient Ham inbox/folder is populated as an ham, otherwise the Spam inbox/folder is populated with the incoming mail as a spam. The threshold is set to 0.5 for decision making to determine if the entire mail is either spam or ham.

Experiment

Extensive empirical evaluation is being performed, using collected large Spam and Ham mail to test the interoperability of the text classifiers; Bayesian Statistical method classifier and the Word Stemming. The investigation includes studies on the effect of the Suspicious terms modification/rearrangement, the introduction of special characters within the terms of local and global sampling, the use of suspicious terms, and the introduction of the Words Stemmer integrated with the Bayesian method to improve the classification accuracy. The data flow diagram of the algorithm is shown in fig. 1. The hybridized algorithm was used because of it classification accuracy and being difficult to outsmart by the scammers advent.

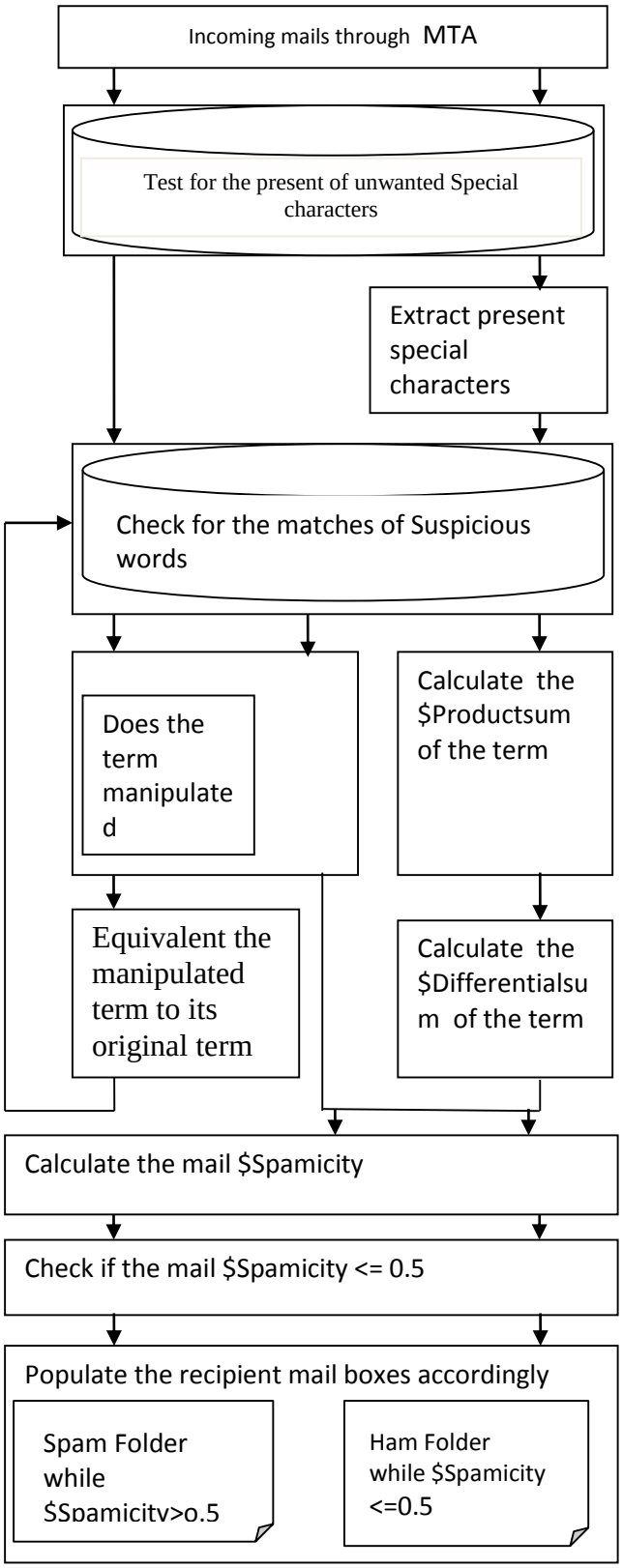

Fig. 1: Algorithm Experiment Data flow diagram.

RESULTS AND DISCUSSION
Result

#

Fig. 2 to 5 are the result of the classification of the algorithm having tested with the known legitimate and malicious mails.

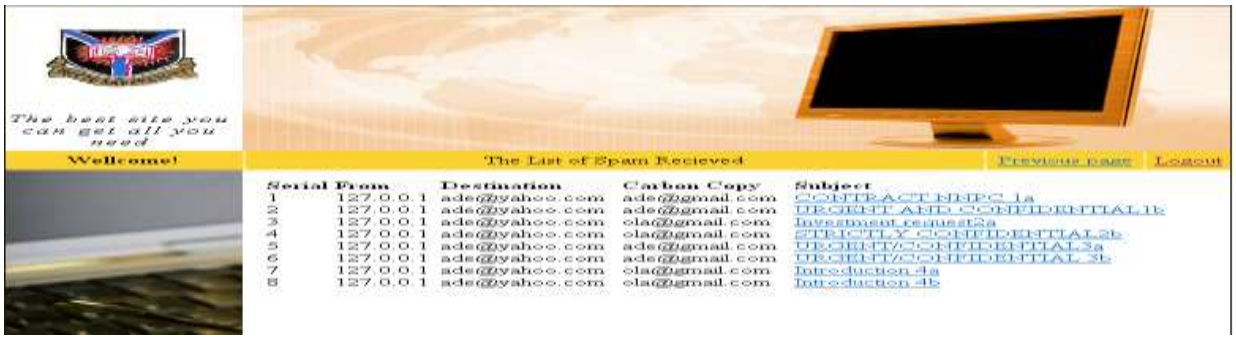

Fig. 2: The list of classified spam mails

Figure 2 is the list of classified spam mails received at the recipient inbox having scan with the Algorithm

Fig. 3: Sample of the Spam Content received

Figure 3 is the **s**ample of the spam mail content received having scan through the mail content with the algorithm, and then classified as a spam.

Fig. 4: The list of Ham mails Received at the recipient inbox.

Figure 4 is the Sample of the Ham mail content received at the recipient inbox having scan through the mail content with the Algorithm, and then classified the mail as Ham.

Fig. 5: Sample of the Ham Content

Table 1 show result of the algorithm tested with known spam mails, those subject names ends with later "a" are spam mails with spam's suspicious content not manipulated and those with the subject names end with later "b" are the spam mails injected with manipulated suspicious words/tokens, before tested with the algorithm. But were all classified as spam

Table 1: Some of the Spam Mails Used for Testing the Hybridized Root Word Extractions Algorithm

S/N	MAIL SUBJECT	IS SUSPICIOUS WORDS/TOKENS MANIPULATED?	KNOWN STATUS	RESULT
1	CONTRACT NNPC 1a	No	Spam	Spam
2	URGENT AND CONFIDENTIAL1b	Yes	Spam	Spam
3	Investment request 2a	No	Spam	Spam
4	STRICTLY CONFIDENTIAL 2b	Yes	Spam	Spam
5	URGENT/CONFIDENTIAL 3a	No	Spam	Spam
6	URGENT/CONFIDENTIAL 3b	Yes	Spam	Spam
7	Introduction 4a	No	Spam	Spam
8	Introduction 4b	Yes	Spam	Spam

Table 2 show result of the algorithm tested with the known legitimates mails and were all classified as malicious

Table 2: Some of the Ham Mails Used for Testing the Hybridized Root Word Extractions Algorithm

S/N	MAIL SUBJECT	KNOWN STATUS	RESULT
1	Festival1	Ham	Ham
2	congratulation	Ham	Ham
3	REBRANDING	Ham	Ham
4	21st CENTURY	Ham	Ham

Table 3: Show result of ordinary Bayesian mail classifier algorithm tested with the known spam mails with and without suspicious terms manipulated, those subject names ends with later "a" are spam mails with suspicious content not manipulated and those with subject names end with later "b" are the spam mails injected with manipulated suspicious words/tokens, before tested with the algorithm. Those with suspicious words manipulated were wrongly classified as ham mail due to the suspicious terms manipulated.

Table 3: Some of the Spam Mails Used for Testing the Ordinary Bayesian Mail Classifier Algorithm

S/N	MAIL SUBJECT	IS SUSPICIOUS WORDS/TOKENS MANIPULATED?	KNOWN STATUS	RESULT
1	CONTRACT NNPC 1a	No	Spam	Spam
2	URGENT AND CONFIDENTIAL1b	Yes	Spam	Ham
3	Investment request 2a	No	Spam	Spam
4	STRICTLY CONFIDENTIAL 2b	Yes	Spam	Ham
5	URGENT/CONFIDENTIAL 3a	No	Spam	Spam
6	URGENT/CONFIDENTIAL 3b	Yes	Spam	Ham
7	Introduction 4a	No	Spam	Spam
8	Introduction 4b	Yes	Spam	Ham

Table 4 show the result of ordinary Bayesian mail classifier against the hybridized root word extractions classifier algorithms, tested with the known spam mail contents with and without suspicious words manipulated. those subject names ends with later "a" are spam mails with suspicious words not manipulated and those with subject names end with later "b" are the spam mails injected with manipulated suspicious words, before tested with the algorithm. Spam mail content with manipulated suspicious words were wrongly classified as ham mails (false negative) using ordinary Bayesian mail classifier. Whereas they were correctly classified as spam mail (true positive) while tested on hybridized root word extractions algorithms.

Table 4: Some of the Spam Mails with and without Suspicious Terms Manipulated Used for Testing the Ordinary Bayesian Mail Classifier against the Hybridized Root Word Extractions Classifier Algorithms

S/N	MAIL SUBJECT	IS SUSPICIOUS WORDS/ TOKENS MANIPULATED?	KNOWN STATUS	RESULT OF ORDINARY BAYESIAN MAIL CLASSIFIER ALGORITHM	RESULT OF HYBRIDIZED ROOT WORD EXTRACTIONS ALGORITHM
1	CONTRACT NNPC 1a	No	Spam	Spam	Spam
2	URGENT AND CONFIDENTIAL1b	Yes	Spam	Ham	Spam
3	Investment request 2a	No	Spam	Spam	Spam
4	STRICTLY CONFIDENTIAL 2b	Yes	Spam	Ham	Spam
5	URGENT/ CONFIDENTIAL 3a	No	Spam	Spam	Spam
6	URGENT/ CONFIDENTIAL 3b	Yes	Spam	Ham	Spam
7	Introduction 4a	No	Spam	Spam	Spam
8	Introduction 4b	Yes	Spam	Ham	Spam

DISCUSSION

Result of the experiment of ordinary Bayesian mail classifier algorithm show that spam mails with manipulated suspicious terms were wrongly classified as ham mail (false negative) due to the suspicious words manipulated. Against the result of the hybridized root word extractions, that combined both Bayesian statistical probability with word stemming algorithm. This effectively identified the suspicious terms that were manipulated to defraud the filter, and classified the spam mails with suspicious terms manipulated as spam (true positive). The classification is correctly and accurately done (true positive and true negative), no false positive or negative is recorded, irrespective of manipulation of the spam mail content. The hybridized algorithm performed effectively on manipulated spam mails as if there was no manipulation. 99% of the tested manipulated spam mails with the algorithm were correctly classified as spam mails and likewise Ham mail were classified correctly as ham mail. However, the algorithm was able to overcome the challenges of false classification as a result of suspicious words manipulations.

CONCLUSION

Spam is one of the major problem internet community is facing today. This paper, explore the effects of statistical Bayesian theorem combined with word stemming filters on suspicious terms manipulated spam mails against the suspicious terms non-manipulated spam mail. The result shows that the algorithm is effective, accurate and not affected by spammer manipulation advent. Results of evaluation of the developed classification models shows that 99% of the suspicious words manipulated spam mails were identified as a spam adding the word stemming features has significantly improved on the ability and capability of the Bayesian classifiers to detect spam emails. However, the algorithm is significantly important in today's internet world.

Direction for Further Research

An interesting research direction, is to devise some sort of dynamic word Stemming algorithm on word boundary detection, where the word boundary is being modified. A good word boundary detection techniques can be used in hybridize with other methods, for more improvement in mail classification. Also time execution variance of hybridized (Bayesian statistical classifier combined with word stemming algorithm) content based filter need to compare and contract against the ordinary Bayesian statistical method in order to identify the execution time.

REFERENCES

Aladdin Knowledge Systems. (2011). Anti-spam white paper.*www.csisoft.com/security/ aladdin/esafe_antispam_whitepaper.pdf*

Albercht. K, (2006). Mastering Spam: A Multifaceted Approach with the Spamato SpamFilter System DSS. ETH NO. 16839

Amol, G. K., Prashant, K. K. and Anil, K. G. (2013). Survey of Spam Filtering Techniques and Tools, and MapReduce with SVM. *International Journal of Computer Science and Mobile Computing.* 2(11): 91 – 98

Anbazhagu, U. V., Praveen J. S., Soundarapandian, R. and Manoharan N. (2014). Efficacious Spam Filtering and Detection in Social Networks. *Indian Journal of Science and Technology*. 7(S7): 180–184. *ISSN (Print): 0974-6846, ISSN (Online): 0974-5645*

Andrej, B., Gordon, V. C., Bogdan, F., Thomas, R. L. and Blaˇz, Z. (2006). Spam Filtering Using Statistical Data Compression Models. Journal of Machine Learning 7:2673-2698. *http://brightmail.com*

Blanzieri, E. and Bryl, A. (2008). A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Revolution*, 29:63–92.

Cranor, L.F. and LaMacchia, B.A. (1998) Spam! *Communications of the ACM (*41): 74-83.http://dx.doi.org/10.1145/280324.280336

Fight Cybercrime (2008) Anti-phishing Techniques SpamAlert.org

Hedieh, S., Golazin, Z. P., Fatemeh, A. (2016). SMS Spam Filtering Using Machine Learning Techniques: A Survey. *Machine Learning Research* 1(1): 1-14. *http://www.science publishinggroup.com/j/mlr* doi: 10.11648/j.mlr.20160101.11

Ja'far, A., Hossam F., Khalid J., Malek A. and Omar A. (2015). Improving Knowledge Based Spam Detection Methods: The Effect of Malicious Related Features in Imbalance Data Distribution. *King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan. Authors and Scientific Research Publishing Inc.* http://creativecommons.org /licenses/by/4.0/

Jonathan A. Z (2003). Classification Data sources. *Spam filter analysis 17.* Com/0101454/stories/2002/09/26/spam Detection.html.

John. W. C (2003). Research Design-Qualitative, Quantitative, and Mixed Methods Approaches. *Sage Publications, Inc., 2nd edition.*

Kanich, C., Weaver, N., McCoy, D., Halvorson, T., Kreibich, C., Levchenko, K., Paxson, V., Voelker, G.M. and Savage, S. (2011). Show Me the Money: Characterizing Spam-Advertised Revenue. *USENIX Security Symposium*, San Francisco.

Mrutyunjaya, P., Ajith, A. and Manas, Ranjan, P. (2011). A Hybrid Intelligent Approach for Network Intrusion Detection. International Conference on Communication Technology and System Design. Procedia Engineering 30 (2012) 1 – 9. *www.sciencedirect.com, www.elsevier.com/locate/procedia*

Nazirova S. (2011). Survey on Spam Filtering Techniques. Communications and Network. 3:153-160 doi:10.4236/cn. *http://www.SciRP.org/journal/cn*

Omar, S., Ashraf, D. and Ramadan, F (2012). A survey of machine learning techniques for Spam filtering. International Journal of Computer Science and Network Security. 12(2):66-73.

Out-law News. (2006). Over 90% of email is spam. Says Spamhaus founder.

Postel. J, (November, 1975). RFC706: On the Junk Mail Problem. Technical report, Network WorkingGroup.

Priyanka, S. and Prashanthi, P. K. (2015). E-mail Spam Classification Using Naïve Bayesian Classifier. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) 4(6): 2792 -2796. ISSN: 2278 – 1323

Reshma, V. and Dhanya, K. A. (2017). Efficient Feature Set for Spam Email Filtering. IEEE 7th International Advance Computing Conference. pp732-737.

Rekha and Sandeep, N. (2014). A Review on Different Spam Detection Approaches. *International Journal of Engineering Trends and Technology (IJETT)*. 11(6): 315-318. ISSN: 2231-5381 *http://www.ijettjournal.org*

Sarah, J. D., Mark, B. and Derek, G. (2012). SMS Spam Filtering: Methods and Data. Dublin Institute of Technology ARROW@DIT Articles School of Computing. pp 1-33

Thamarai, S., Hamid, A. J. and Alaa, Y. T. (2010). Overview of textual anti-spam filtering techniques. *International Journal of the Physical Sciences.* 5(12): 1869-1882.

The Nigerian email Fraud Gallery. (2005). Available online at *www.potifos.com/fraud*

Samir A and Elsagheer M. (2013). Efficient Spam Filtering System Based on Smart Cooperative Subjective and Objective Methods. Int. J. Communications, Network and System Sciences. pp:688-699.

http://dx.doi.org/10.4236/ijcns.2013.62011
http://www.scirp.org/ journal/ijcns

Siham A. M. A., Wadeea A. A. Q., Ahmed K., Ibrahim A. A. A. (2015). Filtering Spam Using Fuzzy Expert System. Journal of Emerging Trends in Computing and Information Sciences**.** 6(12) ISSN 2079-8407 CIS Journal. *http://www.cisjournal.org*

Zhang, L., Zhu, J. and Yao, T. (2004). An evaluation of statistical spam filtering techniques. ACM Transaction on Asian Language Information Process. 3:243–269.